Collocational and colligational patterns in lexical sets: A corpus-based study

Isabel Verdaguer and Anna Poch
University of Barcelona

1. Introduction

The need for a linguistic approach which integrates syntax and semantics in the description of language has now been known for a long time, since the interface between syntax and semantics allows us a systematic and coherent description of the word meaning. The lexicon, from which all types of linguistic information can be projected, thus acquires a relevance which it did not have in the past. A thorough understanding of all the relevant features of lexical items and of their interconnections allow the linguist to study their shared properties and behaviour, establish general patterns and draw generalisations. On the basis of the interdependence between their syntactic properties and their meaning, lexical items can then be classified and grouped into lexical sets.

Following the assumption that the meaning of verbs is interrelated with their syntactic behaviour and that verbs can be classified into lexical sets or classes according to some shared syntactic and semantic properties (Levin 1993), the aim of this study is to see whether verbs forming a lexical set also share similar collocates. In the context of a project[1] dealing with significant collocations in scientific research papers, aimed at helping Spanish scientists to use the right word combinations when writing their papers, we explore what collocations are recurrently used with a selection of verbs which Levin (1993) classifies as *See* verbs and *Sight* verbs. The starting point of this study is thus the question whether verbs which share a similar meaning and similar syntactic behaviour also share similar collocates.

As corpus linguistics studies have confirmed the need to take into account the patterns of word co-occurrence in language description, the importance of collocations and other units of meaning larger than the word has been increasingly acknowledged (Sinclair 1991, 1996, McEnery and Wilson 1996, Gledhill 2000, Stubbs 2002, Oakey 2002). Corpus studies have shown that native speakers do not make up entirely new combinations of words when they speak or write, but very often make use of recurrent chains of words, semi-automatic chunks of language, or multi-word units. Language is therefore a mixture of creativity and repetition. This frequent co-occurrence of words has to be taken very much into account in the description of a language and in lexicography, especially that aimed at non-native speakers of a language, since, although they are usually not difficult to decode, they may cause difficulties in language production. The correct use of words in their context and of the right collocations is essential to show native-like competence and fluency, and it is especially necessary in the writing of scientific papers, which requires a precise expression of one's ideas and of research results. It is important, thus, to know what are the most frequent and expected word combinations used in a specific register and this evidence obviously derives from the observation and analysis of corpora of real text.

Our analysis has started with the selection of a set of verbs which occur in scientific articles. Since observation is a necessary step, crucial in scientific research –most data are first perceived through the sense of sight before being processed by the mind– we have opted for a range of verbs of visual perception: *see, examine, investigate, note, observe* and *view,* which have been classified by Levin (1993) into two different subsets of verbs of perception: *see* is included in the *See* verbs, whereas the rest belong to S*ight* verbs. As we will see, most verbs of physical perception have extended their meanings to refer to mental perception.[2] Some of them, such as *observe*, are verbs mainly of physical perception; in others, such as *investigate,* mental activity is predominant.

[1] This project, reference BFF2001-2988, is financed by the Spanish Ministry of Science and Technology and FEDER

[2] Most of these verbs have their origins as verbs of physical activity and through metaphoric transfer, mainly triggered by abstract direct objects, have become verbs of mental perception as well. Sweetser (1990) has provided ample evidence that change operates in from concrete to abstract domains.

These verbs may be highly polysemous in general language. As Levin's classification is based on the meaning of the verbs, as well as on their syntactic properties,[3] they are obviously found in other groups as well. For example, *see* is also grouped with *Characterize* verbs, along with *view*; *observe* is also classified as a *Conjecture* Verb; *examine, investigate* and *observe* as *Investigate* verbs; *note* and *observe* as *Say* verbs. It can be noticed then that, in addition to sharing the classification of verbs of perception, most of these verbs also share other similar senses and syntactic behaviour which makes it possible to classify them in the same group. This fact also indicates that changes and extensions of meaning follow a similar direction. It is to be expected then, that similar collocates will also be found.

In specialized language, polysemous verbs generally appear in more restricted senses than in general language. Our corpus exploration and analysis aims to find out the patterns of these verbs and their recurrent collocates in scientific texts to see whether generalisations can be reached. We are interested in finding out what verbs can appear with a non-human subject, as well as a personal subject, what type of objects –from a semantic point of view– recur with these verbs, and what generalisations can be drawn from them, or what type of adverbs collocate with these verbs, in addition to their syntactic behaviour. And this can be very useful information in English for Specific Purposes, which will not be found in dictionaries only concerned with specialized vocabulary.

2. Methodology

The following study is based on the analysis of all the occurrences of the verbs *see, observe, note, view, examine and investigate* in a corpus of scientific research articles on biochemistry,[4] which we have compared with a sample of the *British National Corpus*. The software[5] used to process our data has provided the frequency of occurrence and the concordances that we have studied to look for patterns of word co-occurrence.

We have started with the verb *see*, which is a highly frequent verb and may be taken as representative of the verbs of perception.[6] Then we have explored the verbs *observe, view* and *note*, which are verbs more closely related to *look at,* which does not occur in our corpus. Next, we have dealt with *examine*, which could be considered a hyponym of *observe* ("observe attentively") and finally with *investigate.* The verbs of greater frequency of occurrence, *see, observe, examine and investigate,* have been treated separately, whereas the rest have been dealt with in the general discussion. For each verb, the different structures and forms of occurrence were considered to try to set general patterns and at the same time establish differentiations among them.

See verbs "describe the actual perception of some entity" (Levin 1993:186). They take the experiencer or perceiver as subject and what is perceived as object; they can also take *that*-clauses, *-ing* clauses or infinitive clauses as complements;[7] and they do not take the middle alternation. Examples of all these structures can be easily found in a sample of a corpus of general English, such as the *British National Corpus*. However, the analysis of a corpus of English scientific texts provides us with information about the patterns and word co-selection which are specific to this register.

3. Case Studies

3.1. See

In general English, *see* is a very frequent and highly polysemous verb. Some of the meanings found in the *British National Corpus*, however, such as for example "visit" (*We were coming to see you last night*), do not occur at all in our corpus. In scientific English, the meaning of *see* as a verb of visual perception generally implies mental understanding as well. It could be said that there is a continuum and in most occurrences both elements are involved, and the preponderance of physical or mental perception depends on the type of stimulus which co-occurs with it, either a physical entity or a more abstract process or phenomenon. Compare, for example, *no significant staining was seen in the rest of the cell*,

[3] Levin gives special prominence to the expression of arguments and diathesis alternations.
[4] The corpus used in this study consists of 104 articles published in *Biochemical Journal* in the years 1998-1999.
[5] *WordSmith Tools*
[6] The behaviour of *see* has received a lot of attention and has been dealt with in various studies.
[7] Levin does not attempt to deal thoroughly with all the possible complements of *See* verbs.

where physical perception is predominant, and *A similar induction pattern was seen with…* where mental perception is more important. Except for the occurrences of *see* in the imperative form and a very small number of other cases, most occurrences of *see* in which the stimulus or perceived object is a Noun Phrase are passive sentences. And it should be mentioned in this regard that although it is well known that scientific language generally uses the passive voice, whereas with other verbs such as *examine* or *investigate* the Subject *we* is often explicit, with *see*, this is very rare.

Although the perceived object can certainly be expressed by a Noun Phrase whose head is a noun referring to a physical entity such as < *particles, protons or chromosomes* >, the most recurrent collocates are of a more abstract nature, and many of them could be referred to as a state of affairs. Examples are < *effects, trends, presence, changes, levels, pattern, interaction, induction, inhibition …* >, generally either pre- or post-modified to specify their reference (e.g. *A similar pattern of S activation in Z cells is seen / one distinct shift was seen for this fragment*). Past participles with a passive meaning postmodifying similar collocates are also frequent < *the level of induction seen in... / chromosomes frequently seen / the increase seen in cells…*>.

An even more frequent pattern than that of passive structures, however, is the imperative form *see* (60%), which is basically used for references. Its typical collocates are fairly simple Noun Phrases such as: < *table* X, *section* Y, *figure* Z, *appendix* A…>, or adverbs of place: < *below, above…*>. Another frequently recurring pattern is a comment clause introduced by *as* < *As seen in figure, table .../ As can be seen* > as a way of introducing a statement. This is much more frequent than the alternative *that*-clause, of which only two examples have been found, both with the modal *can* (*It can be seen that…*). No examples of *see* followed by an *–ing* clause occur, and only one where passive *see* is followed by a *to*-infinitive clause (*X was seen to…*).

See can also occur in a *to*-infinitive clause, as a predicative complement, or complement to predicative adjectives (*it is possible / difficult to see*) or to nouns (*it is of interest to see*), which mainly convey the speaker's attitude. In 80% of occurrences of *see* in infinitive clauses, the verb is followed by *whether /if* and its meaning then is "find out."

3.2. Observe

The rest of the verbs, which can be classified as *Sight* verbs (Levin 1993:186-187), in general English also take a perceiver as the subject and what is perceived as object, and cannot take an infinitive clause as complement. Some of them (except *observe* or *note*) cannot take a finite *that*-cl, either.[8] And in the same way as *See* verbs, they do not allow the middle alternation.

Observe is the most frequent verb of this group and the one that in principle is most closely connected to physical perception. In general English, *observe* is also highly polysemous. Levin classifies it in four different groups. However, in scientific English, *observe* is mainly used in the sense "look at, watch (attentively)." What can be observed is either something concrete, which is directly perceived, such as: < *proteins, peaks* or *cells* > or, more frequently, something more abstract or a state of affairs, which in many cases needs some sort of mental processing: < *results (similar results…); activity (basal activity, low activity, PLC-bt activity, LPP activity…); processes (phosphorylation, catalysis, dimerization, degeneration…); phenomena (fluorescence, this phenomenon…); interaction (such an interaction…); attachment (some loss of attachment…); effects; trends; characteristics; changes (reduction-sensitive change…).* The Noun Phrases which collocate with *observe* may be again highly complex, with premodification and /or postmodification, and in a a remarkable number of occurrences they indicate differences, similarities or some kind of variation or change: < *A faster migrating X species / An increase in X levels / An accumulatin of PG or Cl mass / Any changes in / Any translocation of / A progressive age-dependent increase* >.

Some sort of comparison is sometimes established: < *A smaller degree of induction was also observed / Similar results are also observed / A similar reduction in / A relatively broad signal is also observed / Increased Ab secretion is also observed / High similarity is also observed* >. Those comparative structures are especially used with the pattern *that / those observed*, where some degree of

[8] Again, Levin does attempt to deal with all the possible complements.

comparison is involved: < *Expression of X was comparable to that observed in /...was greater than that observed in keratinocyte / ...lower than that observed / process comparable to that observed with…*>.

Another interesting pattern is that observed with the past participle. Whereas with all the other verbs, the past participle follows the noun, in the case of *observe*, the past participle can also precede it. And there is a difference in use. The past participle precedes a noun (which can be < *activity / change / curve / discrepancy / kinetics / phenotypes / values / radius / signals / stimulation…*>) when an explanation is provided, and it gives support to what is explained (*the data support the proposal that the observed activation…*), whereas it follows the noun when the writer simply wants to express that something has been observed < *low but significant activities observed / basal activity observed / phosphatase/ luciferase activity observed* >.

Sentences in the active do occur, although there are not many (9,7%). Again the direct objects which are used with *we observed / we have observed* are not physical entities or substances but < *interaction, propensity, increase, difference…*> Or, more frequently, in that case, *observe* is complemented by a finite *that*-clause conveying a statement (*we observed that…*). Interestingly enough, another pattern which occurs with this verb is one indicating negation, which is conveyed by verbs such as *fail* followed by the *to*-infinitive (*we failed to observe*), adjectives such as *unable* (*were unable to observe*) or negative determiners and particles: < *no significant fluorescent staining / no such characteristic staining / no fragmentation / no inhibition was observed* >.

As *see, observe* is used in comment clauses, *as observed / as we have observed*, generally followed by an adjunct of place (*in many other cell types*), or time (*after treatment / previously*). Finite *that*-clauses can occur but are not frequent.

3.3. Examine

Its meaning in scientific texts could be glossed as "to observe attentively in order to discover or find out something" and here again physical perception and mental understanding are involved. Whereas the personal forms of *see* rarely occur in scientific English, *examine* occurs quite frequently with the personal pronoun *we* (26% of the occurrences) and it also allows a Subject which is not human < *Further investigations / those studies / several papers* >. Again, the collocates which appear as direct objects of the active structure do not refer to physical substances or entities, but are more abstract and general, referring to processes, effects, abilities, role, mechanisms, effects, influences, phenomena, role, possibility, interactions …, whereas in the passive structure more concrete entities and substances appear as well < *plates, membranes, granules, cells, nuclei, filters, tissues, specimen, grids ...*>.

Examine may take subordinate clauses introduced by *whether*, and a remarkable proportion of these (40%) follow *examine* in an infinitive clause of purpose (*to examine whether this cell would show…*), which represent 67% of the infinitive clauses occurring with *examine*, the rest occurring with catenatives or as complements to adjectives or nouns which mostly convey the writer's attitude (*it is instructive to examine / it is of interest to examine*). Non-finite *–ing* clauses may also appear in similar structures (*we were interested in examining*). On the other hand, the only modal auxiliary which has been found is *will*, to indicate future time reference. The past participle does not occur often at least in the occurrences found in our corpus, and it is always following the noun (*tissues examined*).

As *examine* may be taken as a step in scientific research, it co-occurs with the adverbs *also* (*we also examined*), *next* (*we next examined*) *then* (*we then examined*) or *further* (*to examine further*), indicating the sequence of events, which can also be conveyed by constructions such as *we began by examining*. It is also for this reason that in the passive sentences, *examined* can appear coordinated with another verb < *membranes were fixed and examined / cells were harvested and examined / starch granules were isolated and examined* >, indicating the different parts of the procedure followed. It is of interest to remark in this respect the presence of the formation *re-examine* (*re-examine the hypothesis / the nature of).*

3.4. Investigate

Whereas in general English, most occurrences of *investigate* are used in the sense "to conduct an enquiry into," which would correspond to Levin's *Investigate* verbs, needless to say, in scientific texts *investigate* is used in the sense "look into from a scientific point of view." The predominant form used

here is the infinitive of purpose, which is again used with the same type of abstract collocates as heads of the Noun Phrase that we have found with the other verbs: < *To (in order to) investigate the effects / the role / the importance / the ability / the function / the mechanism / the distribution / the interaction* >. However, the most significant characteristic which distinguishes it from the other verbs is the fact that it repeatedly co-occurs with expressions indicating possibility < *this possibility / the possible dependence / the possible presence / a possible functional role / this hypothesis* >. Although *examine* can also collocate with *possibility,* the number of such collocations is not as high as with *investigate.* Again, the quest involved in the meaning of this verb is also conveyed by the type of clausal complementation that it takes: interrogative *wh*-clauses < *how / whether / what / which* >. And the especially active implication of the Subject is shown again by the use of the pronoun *we (we investigated / we have investigated)*. Passive sentences are of course also used, and we find here the same pattern that has been previously mentioned in relation to the other verbs: concrete physical objects appear as passive subjects rather than as active direct objects, whereas more abstract or general objects can appear both in the active and passive sentences.

As in the case of *examine,* the adjuncts which are mostly used here are < *further / then / next* > *(we then investigated),* indicating the sequence of events taking place in the investigation, or of manner < *systematically / extensively / directly / thoroughly* >. The need for further work which is involved in the meaning of *investigate*, which has been already indicated by its frequent co-occurrence with the adverb *further* is implied in the structures < *remains to be investigated / will need to be investigated / can be investigated* > or by the formation *un-investigated*. Non-finite clauses with a past participle following the noun can be used but, again like with *examine*, they are not especially frequent in our corpus (2.3%).

4. Discussion and conclusion

Collocations, which were once thought to be an idiosyncratic characteristic of individual words, have now been seen to form general patterns, since the collocates of a node may form sets which are semantically related. This corpus study has revealed the collocational and colligational preferences of verbs belonging to the same lexical set and has shown that they also share general patterns of collocates. However, this study has also revealed that, within these general patterns, each these verbs also have their own preferences, which are closely interrelated with the semantic elements that conform their meaning. We will first comment on the patterns that these verbs share and afterwards we will discuss the particular features that distinguish them.

We will begin with the kinds of perceived objects which occur with the verbs studied. It will come as no surprise that the most frequent structure which occurs in scientific prose is a passive sentence and the perceived object will consequently be a passive subject. The passive is predominatly used with all the verbs in this register, except with *investigate*, whose to-infinitives (of purpose) are the most frequent forms, and *see* when it is used for references. Except with the verb *view*, which has a physical object < *cells* > when it is used as a *Sight* verb, the head of the complex Noun Phrases that conform the passive subject are not predominantly nouns referring to physical substances or objects, as could be expected when dealing with verbs of perception, but more abstract entities, such as effects, results, processes or phenomena, which require some mental processing. This indicates that all the verbs sharing this feature also share the extension of meaning from physical to mental perception. Unlike the active structure, however, which is typically used with abstract direct objects, except in the use of *see* mentioned above, in the passive structure we can also find physical entities, such as *cells, granules* or *images* which can be directly perceived. The fact that in the active voice we have only found abstract entities or state of affairs[9] can be explained by the explicit involvement of the perceiver and the cognitive process which is implied here, which is not necessary in the direct perception of a physical entity, and thus the perceiver can be left implicit. Within this general pattern, the idiosyncratic characteristics of some verbs have to be mentioned: *observe,* and also *note,* frequently co-occur with Noun Phrases indicating similarities, differences, increase, decrease or some variation or change. *Investigate*, on the other hand repeatedly collocates with Noun Phrases indicating possibility.

The perceiver is in most cases left implicit, as corresponds to passive structures. However, here again there are verbs which stand out, since *observe, examine* and *investigate* have a fairly high percentage of occurrences in active sentences with the explicit Subject *we*. In the case of *observe*, *we*

[9] Except with imperative *see*.

observed is frequently followed by a *that* subordinate clause, introducing a statement or observation (*we observed that*). On the other hand, *examine* allows a metonymic extension of the Subject, which can be < *studies / papers / investigation…*>.

The imperative form occurs with two verbs *see* and *note*, but with some differences in use and in collocates. In both cases, they are used to draw attention to something, but whereas *see* collocates with nouns such as *tables* or *figures,* for example, *note* collocates with lexical items which refer to properties or processes (*note the co-precipitation of / the distribution / the co-enrichment of)* or is followed by a statement introduced by a *that*-clause.

Observe and *see,* as well as *note,* may take a subordinate content clause introduced by *that*, conveying a fact or an idea. The same verbs may also take a comment clause introduced by *as* (much more frequently in the case of *see)* typically co-occurring in this structure with an adjunct. They usually collocate with adjuncts of place (*as is seen in Figure X* / *as observed in many other cell type*), of manner (*as can be seen by comparing…),* of time (*as observed / noted previously)*. Note in this last example, that the meaning of both verbs in this context is "make a comment or remark", which is also the case when *note* is followed by the agent (*As noted by X)*. *Examine, investigate* and *see* (in the sense "find out"), on the other hand, can take a *wh*- interrogative clause.

Past participles, with passive meaning, usually occur following the head of the Noun Phrase, (that is to say they occur postpositively). And in this structure, the meaning of *observe* occurring with *previously (species previously observed)* is of physical perception. *Observed* also stands out in the corpus because, differently from the other verbs, whose past participle follows the noun, *observed* can precede it to indicate not simply that something has been observed, but to provide an explanation for something.

We have found examples of *to*-infinitives as predicative complements (*the aim/ purpose of this study was to investigate)* or as complements to adjectives (*it was of interest to investigate*), nouns (*it was difficult to see / it is possible to see/ it is important to note*), or verbs (*we would expect to observe)* which mostly express some kind of modality or the speaker's attitude. The modal auxiliary *can*, which typically occurs with *see* and other verbs of perception, is also recurrently found in the corpus. Other linguistic expressions of possibility, illustrated by Noun Phrases such as *this possibility* are used with *examine* and especially with *investigate*. These two verbs are also used with an infinitive of purpose, again indicating the search for new discoveries involved in the meaning of the verb.

Apart from the adjuncts, such as those of place, time or manner, which are generally used, we also find collocational preferences which are related to the specific semantic elements of the different verbs. Along with the adjuncts indicating the sequence of events occurring with *examine* or *investigate*, we have found coordinated verbs *(fixed and examined*) or occurrences of constructions expressions such as *remains to be investigated*.

In conclusion, verbs which have similar syntactic and semantic properties have been shown to share similar collocational patterns, too. Although specific differences within general patterns may arise, the more properties verbs have in common, the more similar their collocational preferences are.

References

Gledhill C 2000 The discourse function of collocation in research article introductions. *English for Spacific Purposes.* 19 (2): 115-135.

Levin B 1993 *English verb classes and alternations.* Chicago and London, The University of Chicago Press.

McEnery A, Wilson A 1996 *Corpus linguistics*. Edinburgh, Edinburgh University Press.

Oakey D 2002 Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In Reppen R, Fitzmaurice S, Biber D (eds), *Using corpora to explore linguistic variation*. Amsterdam / Philadelphia, John Benjamins Publishing Company.

Sinclair J 1991 *Corpus, Concordance, Collocation*. Oxford, Oxford University Press.

Sinclair J 1996 The search for units of meaning. *Textus* 9: 75-106.

Sweetser E 1990 *From etymology to pragmatics*. *Metaphorical and cultural aspects of semantic structure.* Cambridge, Cambridge University Press.

Stubbs M 2002 *Words and Phrases*. Oxford, Blackwell.